Chain Rules for Entropy

The entropy of a collection of random variables is the sum of conditional entropies.

Theorem: Let X_1, X_2,…X_n be random variables having the mass probability $p(x_1,x_2,\ldots x_n)$. Then

$$H(X_1, X_2, \ldots X_n) = \sum_{i=1}^{n} H(X_i | X_{i-1}, \ldots X_1)$$

The proof is obtained by repeating the application of the two-variable expansion rule for entropies.

Conditional Mutual Information

We define the conditional mutual information of random variable X and Y given Z as:

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z)$$
$$= E_{p(x,y,z)} \log \frac{p(X,Y|Z)}{p(X|Z)\,p(Y|Z)}$$

Mutual information also satisfy a chain rule:

$$I(X_1, X_2, \ldots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y | X_{i-1}, X_{i-2}, \ldots X_1)$$

Convex Function

We recall the definition of convex function.

A function is said to be *convex* over an interval (a,b) if for every x1, x2 ∈(a.b) and 0≤ λ ≤ 1,

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2)$$

A function f is said to be *strictly convex* if equality holds only if λ=0 or λ=1.

Theorem: If the function f has a second derivative which is non-negative (positive) everywhere, then the function is convex (strictly convex).

Jensen's Inequality

If f is a convex function and X is a random variable, then

$$Ef(X) \geq f(EX)$$

Moreover, if f is strictly convex, then equality implies that X=EX with probability 1, i.e. X is a constant.

Information Inequality

Theorem: Let p(x), q(x), x ∈χ, be two probability mass function. Then

$$D(p\|q) \geq 0$$

With equality if and only if

$$p(x) = q(x) \qquad \text{for all x.}$$

Corollary: (Non negativity of mutual information): For any two random variables, X, Y,

$$I(X;Y) \geq 0$$

With equality f and only if X and Y are independent

Bounded Entropy

We show that the uniform distribution over the range χ is the maximum entropy distribution over this range. It follows that any random variable with this range has an entropy no greater than log|χ|.

Theorem: H(X)≤ log|χ|, where|χ| denotes the number of elements in the range of X, with equality if and only if X has a uniform distribution over χ.

Proof: Let u(x) = 1/|χ| be the uniform probability mass function over χ and let p(x) be the probability mass function for X. Then

$$D(p\|q) = \sum p(x)\log\frac{p(x)}{u(x)} = \log|\chi| - H(X)$$

Hence by the non-negativity of the relative entropy,

$$0 \leq D(p\|u) = \log|\chi| - H(X)$$

Conditioning Reduces Entropy

Theorem:

$$H(X|Y) \le H(X)$$

with equality if and only if X and Y are independent.

Proof:

$$0 \le I(X;Y) = H(X) - H(X|Y)$$

Intuitively, the theorem says that knowing another random variable Y can only reduce the uncertainty in X. Note that this is true only on the average. Specifically, H(X|Y=y) may be greater than or less than or equal to H(X), but on the average

$$H(X|Y) = \sum_y p(y)H(X|Y=y) \le H(X)$$

Example

Let (X,Y) have the following joint distribution

Y \ X	1	2
1	0	3/4
2	1/8	1/8

Then H(X)=(1/8, 7/8)=0,544 bits, H(X|Y=1)=0 bits and H(X|Y=2)=1 bit. We calculate H(X|Y)=3/4 H(X|Y=1)+1/4 H(X|Y=2)=0.25 bits. Thus the uncertainty in X is increased if Y=2 is observed and decreased if Y=1 is observed, but uncertainty decreases on the average.

Independence Bound on Entropy

Let $X_1, X_2, \ldots X_n$ are random variables with mass probability $p(x_1, x_2, \ldots x_n)$. Then:

$$H(X_1, X_2, \ldots X_n) \leq \sum_{i=1}^{n} H(X_i)$$

With equality if and only if the X_i are independent.

Proof: By the chain rule of entropies:

$$H(X_1, X_2, \ldots X_n) = \sum_{i=1}^{n} H(X_i | X_{i-1}, \ldots X_1) \leq \sum_{i=1}^{n} H(X_i)$$

Where the inequality follows directly from the previous theorem. We have equality if and only if X_i is independent of $X_1, X_2, \ldots X_n$ for all i, i.e. if and only if the X_i's are independent.

Fano's Inequality

Suppose that we know a random variable *Y* and we wish to guess the value of a correlated random variable *X*. Fano's inequality relates the probability of error in guessing the random variable *X* to its conditional entropy $H(X|Y)$. It will be crucial in proving the converse to Shannon's channel capacity theorem. We know that the conditional entropy of a random variable *X* given another random variable *Y* is zero if and only if *X* is a function of *Y*. Hence we can estimate *X* from *Y* with zero probability of error if and only if $H(X|Y) = 0$.

Extending this argument, we expect to be able to estimate *X* with a low probability of error only if the conditional entropy $H(X|Y)$ is small. Fano's inequality quantifies this idea. Suppose that we wish to estimate a random variable *X* with a distribution $p(x)$. We observe a random variable *Y* that is related to *X* by the conditional distribution $p(y|x)$.

Fano's Inequality

From *Y*, we calculate a function *g(Y)* = *X^* , where *X^* is an estimate of *X* and takes on values in *X^*. We will not restrict the alphabet *X^* to be equal to *X*, and we will also allow the function *g(Y)* to be random. We wish to bound the probability that *X^* ≠ *X*. We observe that *X* → *Y* → *X^* forms a Markov chain. Define the probability of error: *Pe* = Pr{*X^* = *X*}.

Theorem:

$$H(P_e) + P_e \log(|\chi| - 1) \geq H(X \mid Y)$$

$$1 + P_e \log|\chi| \geq H(X \mid Y)$$

The inequality can be weakened to:

$$P_e \geq \frac{H(X \mid Y) - 1}{\log|\chi|}$$

Remark: Note that $P_e = 0$ implies that H(X|Y) = 0 as intuition suggests.